Public Choice (2020) 185:359–376
https://doi.org/10.1007/s11127-019-00694-4

Causal inference and American political development: contrasts and complementarities

Devin Caughey[1] · Sara Chatfield[2]

Received: 10 July 2019 / Accepted: 17 July 2019 / Published online: 23 July 2019

Abstract

Causal inference and American political development (APD) are widely separated and (to some) fundamentally incompatible tendencies within political science. In this paper, we explore points of connection between those two perspectives, while also highlighting differences that are not so easily bridged. We stress that both causal inference and APD are centrally interested in questions of causation, but they approach causation with very different ontological and epistemological commitments. We emphasize how the sort of detailed, contextualized, and often qualitative knowledge privileged by APD can promote credible causal (and descriptive) inferences, but also that scholars of causal inference can benefit from alternate conceptions of causality embraced by APD work. We illustrate with two empirical examples from our own research: devising weights for quota-sampled opinion polls and estimating the political effects of the Tennessee Valley Authority. We conclude that bringing APD and causal inference together on more equal terms may require a broader perspective on causation than is typical of scholarship in the causal-inference tradition.

Keywords Causal inference · American political development · Survey research · Policy feedback

JEL Classification N4 · C0 · H4

✉ Sara Chatfield
Sara.Chatfield@du.edu

Devin Caughey
caughey@mit.edu

[1] Department of Political Science, Massachusetts Institute of Technology, E53-463, Cambridge, MA 02139, USA

[2] Department of Political Science, University of Denver, Sturm Hall, Room 473, 2000 E. Asbury Ave., Denver, CO 80208, USA

1 Introduction

Although their emergence as self-conscious movements in political science occurred roughly contemporaneously, causal inference (CI) and American political development (APD) have evolved largely in isolation from, if not in active opposition to, one another. In the 1970s and 1980s, while the statistician Donald Rubin was developing concepts and methods of CI later imported into political science (see, e.g., Sekhon 2008), APD was emerging as an "insurgent movement" led by Stephen Skowronek, Karen Orren, and other social scientists who sought to reinvigorate the historical study of American political institutions (Mettler and Valelly 2016, p. 1). As this special issue attests, many scholars have sought to keep a foot in both camps. The theoretical commitments and empirical practices of the two are so disparate, however, that maintaining that dual footing requires considerable agility. Our goal in this paper is to bridge the gap between APD and CI by exploring points of connection and synergy while also highlighting and explaining differences that are not bridged so easily.

Echoing the advice of Dunning (2012) and others, we emphasize the ways that credible design-based causal inferences depend on the sort of detailed, often qualitative knowledge of a particular context that APD privileges. Such knowledge is necessary not only to validate the identification assumptions of a given research design, but also to interpret the substantive meaning of causal estimates. At the same time, however, we also argue that the kinds of causal arguments typical of APD, with its focus on macropolitical outcomes and interactions among path-dependent historical processes, are difficult to couch in terms of CI, which focuses on the average effects of discrete, manipulable "treatments". We suggest that alternative perspectives on causation—ones that take a more structural, regularity-based approach and that recognize the value of mixed-method evidence on causal processes—may be a better fit for much APD research.

We illustrate the foregoing themes with examples drawn from two of our own projects. The first is our work on constructing adjustment weights for quota-sampled public opinion polls from the 1930s to 1950s (Caughey et al. 2017). That project highlights two points: the essential role of qualitative knowledge in understanding and analyzing imperfect quantitative data and the importance of valid descriptive inference for drawing causal conclusions. The second example is our research on the political feedbacks of the Tennessee Valley Authority (TVA; Caughey and Chatfield 2016). Like the first example, our study of the TVA highlights the importance of historical knowledge for causal inference, but it also shows how strong causal designs can provide quantitative evidence for such core APD concepts as policy feedback. However, the second project also illustrates the limits of nonparametric CI in the Rubin tradition and the utility of model-based (structural) approaches to causation.

We conclude the paper with a plea for greater engagement and mutual respect between APD and CI. The two fields have much to learn from each other and have more overlap than is sometimes recognized. At the same time, we caution that such engagement should not take the form of subsuming one perspective within the other. In particular, APD should not be made a mere handmaiden of causal inference, serving only as a historical storehouse of natural experiments and a source of evidence in support of identification assumptions. Rather, causally oriented scholars should recognize that CI is not the only way—or even the only quantitative way—of conceptualizing and studying causation. Enlarging the tent of causal inference to include both model-based quantitative approaches and rigorous and well-reasoned qualitative research will expand CI's substantive scope and its integration with alternative research traditions. And, for APD scholars, deep knowledge of historical data can offer opportunities for compelling research designs in the CI-style that can complement qualitative evidence and act as a building block in telling broader causal stories.

2 Conceptual background

2.1 Causal inference

Causation is the relation between two events, one of which (the "cause") produces or brings about the other (the "effect"). In a general sense, causal inference is the drawing of conclusions about particular instances of causation. Within political science, however, the term *causal inference* also is used to refer to a specific approach and scholarly tradition associated primarily with the statistician Donald Rubin. For the remainder of this article, we will use the abbreviation *CI* to refer to this more specific meaning and *causal inference* to the more general one.

Sometimes referred to as the Rubin causal model (RCM), CI in this more specific sense has two basic elements. The first is its *manipulationist* view of causation: Only events ("treatments") that could, in principle, be manipulated in a hypothetical experiment can count as causes ("no causation without manipulation"; Holland 1986, p. 959). That element generally is interpreted as ruling out as causes attributes, such as sex or race, for which it is hard to imagine, let alone implement, well-defined manipulations (but see Sen and Wasow 2016).

The second core component of the RCM is its definition of causal effects in terms of *counterfactuals*: The effect of an observed cause is the difference between the observed outcome and the outcome that would have been observed had, counter to fact, the cause been absent (Neyman 1923). In formal notation, the causal effect for unit i is $\tau_i = Y_i(1) - Y_i(0)$, where $Y_i(d)$ is the "potential outcome" with the cause D_i set to $d \in \{0, 1\}$. Of course, in reality, an individual cannot both receive and not receive treatment simultaneously. Because the causal effect is defined at the level of the unit i (where a "unit" is an opportunity to apply or withhold treatment), treatment effects never can be observed directly, which is what Holland (1986) calls "the fundamental problem of causal inference" (p. 947). Instead, researchers must use a careful research design to make inferences about the distribution of τ_i across units. For example, in experimental work, a researcher might assign treatment randomly to ensure that individuals in the treatment and control groups are otherwise similar. That focus on the *assignment mechanism*—the process by which units are either exposed or not exposed to the treatment—is a signal feature of the RCM.

One of the foundational insights of the RCM is the importance of distinguishing clearly between ontology (causal effects) and epistemology (causal inference). That is, the RCM defines the objects of scientific interest—the individual-level causal effects τ_i, which fundamentally are unobservable—independently of the methods and assumptions—in particular, the assignment mechanism—used to justify inferences about these effects (Rubin 2010, p. 43). In that sense, the RCM is a model of causation as well as causal inference.

Nevertheless, the view of causation provided by the RCM is quite spare. The fundamental building blocks of the RCM are individual-level causal effects. Subject to logical constraints (e.g., that treatment cannot have had a negative effect on a treated unit whose observed outcome is the maximum achievable), the individual-level effects may be arbitrarily heterogeneous. Causal effects are summarized by averaging over their distribution in the sample, and they are generalized by averaging over the population from which units putatively were sampled (e.g., Hartman et al. 2015). Effects may differ on average across subpopulations, but moderating variables are not considered to have *caused* that variation in effects unless the moderator itself is a well-defined and identified treatment.

More fundamental forces, regularities, or covering laws—ideas central to many philosophical conceptions of causation (Brady 2009; Paul and Hall 2013)—have no dedicated place in the RCM, at least as practiced in political science.[1] Nor do causal mechanisms—the intermediate processes through which causes bring about their effects—play formal roles, although under certain conditions it may be possible to estimate the effects of mediating variables (Imai et al. 2011). The RCM's neglect of those aspects of causation can make it awkward to draw inferences about general theories or models from the statistical results of a particular application of the RCM. Such difficulties are particularly salient in APD scholarship, which often focuses on a single, non-generalizable case developing over time, the interaction of multiple causal forces or factors, or both.

It thus bears emphasizing that the RCM is not the only statistical framework for causal inference. The most salient alternative approach is Judea Pearl's (2000) Structural Causal Model (SCM). Unlike Rubin's, Pearl's approach rests fundamentally on structural models of causal relationships between variables, which are represented visually using causal graphs. The SCM subsumes the RCM in the sense that the unit-level counterfactual quantities (i.e., potential outcomes) that are the basic building blocks of the RCM can be derived from causal graphs (Pearl 2009, pp. 126–32). An advantage of Pearl's graphical approach is that it highlights the possibility of multiple strategies for identifying the same causal quantity. In particular, it sometimes is plausible to achieve identification either by balancing determinants of treatment, as emphasized by the RCM, or by adjusting for alternative causes of the outcome (Morgan and Winship 2015, pp. 128–30). In addition, the close connection in the SCM between general theory and specific causal quantities often make it a particularly appealing alternative or complement to the RCM in empirical APD studies.[2]

2.2 American political development

American political development (APD) as a subfield of political science became a force in the 1980s with a renewed emphasis on the importance of careful attention to a historical understanding of American politics. The term itself was popularized by Karen Orren and Stephen Skowronek's journal *Studies in American Political Development*, launched in 1986. The APD movement grew out of historical institutionalism, which aimed primarily to attend to the role of institutions—social, political, and economic—in ordering and explaining a variety of political phenomena in the United States and elsewhere in the world (March and Olsen 1984). As distinct from rational choice approaches that emphasize institutional equilibria, historical institutionalism tends to focus on the historical embeddedness

[1] Rubin has sought to address the problem of generalization in the absence of random sampling by using Bayesian predictive models for the potential outcomes, which he considers the "third leg of the RCM" (Rubin 2010, p. 45). To our knowledge, however, that approach rarely if ever has been implemented by political scientists.

[2] A third influential perspective on causal inference is that associated with the social psychologist Donald Campbell (Campbell and Stanley 1963; Shadish 2010). Like Rubin, Campbell is centrally concerned with research design as a basis for causal inference. More than Rubin, however, Campbell focuses on specific threats to causal inference ("internal validity") and on generalizing causal effects ("external validity").

of institutions and the processes by which they are created and transform over time (Thelen 1999). APD scholarship is diverse and wide-ranging, but generally tries to move beyond the use of historical examples or anecdotes to the development of broader theories of institutional and governance changes in U.S. politics.

The central concerns of APD include the role of the state, formal and informal institutions, and ideas and culture, all set within rich historical context. Mettler and Valelly (2016) call APD a "wide-angle lens" on politics, in which scholars primarily are focused on macro-level political processes, systematic structural changes, and broad forces that operate over time. Much, though certainly not all, APD work focuses on research designs examining essentially an *N* of 1: "America." As such, APD's view of causality can differ quite dramatically from that found in other approaches to political science.

Take, for example, the concepts of path dependence, critical junctures, and policy feedback. Theories of path dependence emphasize that a range of outcomes often are possible given the same set of conditions at an initial moment of causal openness (a "critical juncture"). Through self-reinforcing mechanisms that raise the cost of shifting to a new path as time passes, small, random events early in a political process can have much larger consequences later in time. Sequencing and temporality thus are enormously important when path dependence is at play. A given "cause" may have dramatically different effects depending on whether it occurs early or late in the process (Pierson 2000). Path dependence is often critical to analyses of policy feedback, which consider the ways that government policies affect the larger political system, changing the possibilities for future policymaking (e.g., Pierson 1993; Hacker 2002). Those sort of reciprocal causal processes are hard to represent in terms of a singular treatment assigned at a specific moment in time.

APD research tends to emphasize interactions among multiple causes. Orren and Skowronek (2004) suggest the term *intercurrence* to describe the contemporaneous operation of multiple, conflicting political orders, leading to tension and, ultimately, change. Similarly, *layering*, a phenomenon defined and developed by Schickler (2001), occurs when preexisting institutions are difficult to change or eliminate, and so new ones are layered on top, causing institutions to build up over time in often contradictory ways. In both of those cases, a single "cause" is insufficient to explain an outcome. Instead, causality here relies on the particular, often contingent ways in which causes interact and collide in a given historical setting. Each of the approaches common to APD scholarship can make for uneasy fits with CI-style analysis.

In following two sections, we discuss two of our own research projects that in different ways address the intersection of causal inference and American political development. The first project involves the construction of weights for quota-sampled opinion polls from the 1930s–1950s, and the second the policy feedbacks of the Tennessee Valley Authority.

3 Example 1: Weighting quota-sampled opinion polls

By the time the first nationally representative academic survey of American political attitudes, the 1952 American National Election Study (ANES), was conducted, national opinion polling was already a decade and a half old. Between 1935 and 1952, George Gallup, Elmo Roper, and other commercial pollsters had administered hundreds of such surveys (Converse 1987). Thanks to the efforts of the Roper Center for Public Opinion Research (https://ropercenter.cornell.edu), the individual-level data from many of those polls have been archived and made available to the public. Until recently, however, academic researchers by and large neglected that rich data source, for two main reasons. First, the data files were difficult to analyze in the raw, uncoded form in which they were archived. Second, scholars were unsure of the representativeness of the polling samples, most of which were gathered using quota-controlled sampling techniques rather than the probability-based methods of the ANES and other later surveys (Berinsky 2006). To remedy that neglect, Adam Berinsky and Eric Schickler have led a collaborative effort to clean and code the data files and to develop statistical methods for analyzing them (Berinsky et al. 2011).[3] Our primary involvement with this project has been in the creation of survey weights designed to render the polling samples more representative of the U.S. public (Caughey et al. 2017). In this section, we use the project to illustrate some of the ways that APD and CI can complement one another.

Given that early opinion polls provide crucial insights into the U.S. public during several important episodes in American political history—including the Great Depression, the New Deal, the Second World War, the early Cold War, and the first stirrings of the Civil Rights Movement—the project's connection to American political development is clear. Less obvious, perhaps, is its connection to causal inference, given that the primary purpose of polling samples is descriptive inference regarding the distribution of opinion in the population. To such skepticism, we offer two rejoinders. First, valid causal inference relies on valid descriptive inference; we cannot estimate accurately the causal relationships between variables that are not well measured. Second, and more specific to our particular application, the effectiveness of survey weights hinges on a detailed understanding of what *caused* some respondents' answers to be recorded and not others', as well as an understanding of the determinants of citizens' attitudes. That is, effective survey weighting depends on having good models of the causal processes underlying the sampling process, the nonresponse mechanism, and the outcomes of interest.

3.1 Adjusting unrepresentative samples for better descriptive inferences

Before 1948, nearly all opinion polls employed a combination of area sampling of interview locations and quota-controlled sampling of individual respondents. After apportioning the sample (usually 2000–3000 respondents) across geographic regions, polling organizations purposively selected interview locations (e.g., cities, towns, or farm districts), generally aiming to select a mix of locales politically representative of the state as a whole. At each site, interviewers were instructed to select a range of respondents, subject to detailed demographic quotas. Some of those quotas, such as for gender, were strict, but for others, such as age and economic class, interviewers merely were instructed to get "a good spread" (Berinsky 2006, p. 504).

Though remarkably accurate for certain purposes, especially predicting election outcomes, quota-sampled polls were vulnerable to two main sources of bias. First, because pollsters often were most interested in election outcomes, they created quotas intended to be representative of the electorate rather than of the adult population. Most egregiously, many polls failed to include a single African American respondent from the U.S. South, where nearly all blacks effectively were disenfranchised. Less extreme imbalances likewise

[3] Berinsky et al. (2011) and Caughey et al. (2017) have been funded by two National Science Foundation grants: SES-0550431 (Adam Berinsky and Eric Schickler, "Collaborative Research: The American Mass Public in the 1930s and 1940s," 2006–2010) and SES-1155143 (Adam Berinsky, Eric Schickler, and Jasjeet S. Sekhon, "Collaborative Research: The American Mass Public in the Early Cold War Years," 2012–2014).

were evident in the distribution of other demographic attributes, such as gender and region. Second, in addition to such intentional departures from representativeness, the discretion of interviewers in selecting respondents (and of potential respondents in deciding whether to take the survey) introduced further discrepancies within quota-controlled categories. In particular, markers of economic class and educational attainment, which were not explicitly controlled, tended to be much larger in the poll samples than in the U.S. population as a whole. As a consequence, political attributes correlated with class status, such as Republican Party identification, are similarly over-represented in the sample relative to the population.

Whether intentional or unintentional, the discrepancies introduced the potential for bias in the estimation of population quantities. In general, nonresponse bias is a function of the population correlation between the outcome of interest (y) and the probability of responding (ρ). Nonresponse bias can be reduced, if not fully eliminated, by weighting respondents such that the correlation between y and ρ in the weighted sample is as small as possible. A general framework for constructing such adjustment weights is calibration (Deville and Särndal 1992), which entails finding a set of weights that satisfy a set of K moment constraints:

$$t_k = \sum_i w_i z_{ik},\ k \in 1 \dots K,$$

where the *auxiliary vector* z_{ik} is a function of one or more *auxiliary variables* measured in both the sample and population. Special cases of calibration include raking, which matches the marginal population distributions of a set of categorical auxiliary variables, and post-stratification, which matches the variables' joint distribution. Both of the two main tasks of calibration—selecting the auxiliary vector and constructing population targets—require context-specific substantive knowledge.

The choice of auxiliary vector $\mathbf{z}_i$ is crucial because the reduction in nonresponse bias depends on how well $\mathbf{z}_i$ predicts y_i and ρ_i. If either y_i or ρ_i is a linear combination of $\mathbf{z}_i$, then calibration estimators are consistent even in the face of nonresponse (Särndal and Lundstrom 2005). In other words, effective calibration requires a good model of the determinants of nonresponse and/or the outcome of interest (preferably both). Because of data sparsity, it often is not feasible to weight samples to match the joint distribution of all auxiliary variables, so substantive judgement must be used to select the most important variables and their interactions that will be included in $\mathbf{z}_i$. In our case, for example, we knew that individuals who were male, white, non-Southern, educated, and wealthy—all correlates of political attitudes—were more likely to be sampled, so we made sure to include indicators for those attributes in the auxiliary vector whenever possible. More subtly, we also knew that the undersampling of women was less severe in the South than outside it but that the gender gap in voter turnout was larger. That is, both y and ρ depended on the interaction of gender and region, meaning that their interaction also should be included in the auxiliary vector.[4]

Though typically neglected in survey weighting texts, constructing population targets often is far from straightforward. In our application, the potential auxiliary variables are *State*, *Black*, *Female*, *Education*, *Professional*, *Phone Ownership*, *Farm*, *Urban*, and *Age*.

[4] The severe undersampling of Southern (but not non-Southern) African Americans sometimes requires the even more drastic response of redefining the target population to what we call the "voting-eligible population" (i.e., white Southerners plus all non-Southerners).

For some of those variables, merely ensuring that auxiliary variables are measured comparably in the sample and population can be a labor- and information-intensive task. Take, for example, the seemingly simple construct of "phone ownership." From records obtained from AT&T's corporate archives, we derived data on the number of residential telephones in each state in several years (1930, 1935, 1937, 1940, and 1945). To convert those numbers into phone-ownership rates, we had to adjust for the fact that some households had two lines while some lines were shared between households. We then had to adjust for the fact that early opinion polls, although they usually asked about phone access, employed ten different question variants at different points in time (e.g., "Is there a telephone in your home?" versus "Is your home telephone number in the telephone book?"), only one of which corresponds to the ownership rate derived from the AT&T data. Reconciling the discrepancies required statistical adjustments informed by a good deal of historical detective work (see Caughey et al. 2017, Appendix A.2).

Additional complications were introduced by the fact that the population data on auxiliary variables was from multiple distinct sources available irregularly over time. Thus, to estimate variables' joint distribution at each point in time, we had to develop a dynamic ecological inference model (Caughey and Wang 2019). As noted above, information on the auxiliary variables' joint (as opposed to marginal) distributions is important because we know that both the probability of response and important political attitudes depended on the interaction of different variables. That the targets be dynamic is equally important, for the demographic composition of the United States changed markedly over the period. For instance, the Great Migration led to significant changes in the joint distribution of *Black* and *Region*. In 1936, only 33% of African Americans lived outside the former Confederacy, but by 1952 fully 44% did.

In short, constructing calibration weights for quota-sampled polls requires the close integration of statistical analysis and substantive knowledge, that attends to causal processes in survey sampling and response. Although it is unlikely that the weights eliminate all nonresponse bias, they do succeed in improving the polls' accuracy with respect to out-of-sample benchmarks (Caughey et al. 2017, pp. 17–21). Adjusting for nonresponse bias can in turn yield materially different estimates of public opinion, as we show in the following section.

3.2 Implications for the New Deal realignment

Though not themselves causal quantities, descriptive estimates of public opinion can be essential ingredients in causal inference. Consider, for example, the theory of electoral realignments, one of the most influential perspectives on American political development (Burnham 1967; Sundquist 1983). A central claim of realignment scholars is that the 1930s marked a transition from the Republican-dominated "system of 1896" to a "New Deal system" in which Democrats enjoyed a popular majority for at least a generation. Despite broad agreement on the existence of that realignment,[5] scholars disagree over its causes. The traditional view is that the Great Depression and the Democrats' policy responses to it caused a durable shift in the parties' coalitional bases and relative support in the mass

[5] Even Mayhew (2002), who generally is critical of realignment theory, finds that the New Deal realignment holds up better than other putative cases of partisan realignment.

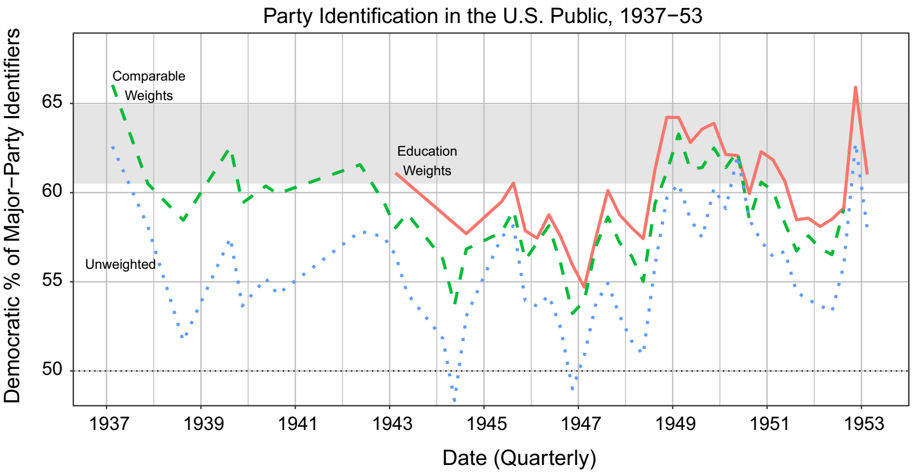


Fig. 1 Weighted and unweighted trends in party identification in the U.S. public, 1937–1953. The dotted line tracks the unweighted quarterly percentages of Democrats among major-party identifiers. The dashed and solid lines track the same quantity estimated using comparable and education weights, respectively. The horizontal gray band represents the 95% confidence interval around the average Democratic percentage recorded in ANES surveys of the 1950s. Note that comparable weights are not available after the third quarter of 1952

public (e.g., Key 1964; Ladd and Hadley 1975; Sundquist 1983).[6] An alternative view, offered by Norpoth et al. (2013), argues that the true cause of the Democratic Party's postwar majority was not the Depression and New Deal of the 1930s, but rather the party's successful management of war and the economy during the 1940s.

Norpoth et al. (2013) support that argument by making innovative use of nearly 200 quota-sampled opinion polls conducted in that period. The rich data source enables them to construct time series of Democratic and Republican party identification between 1937 and 1952. They find that it was not until 1948 that Democrats gained the commanding advantage in the mass public recorded in ANES surveys of 1952 and after (Norpoth et al. 2013, 147). They adduce this as evidence in support of their claim for the causal primacy of Democrats' management of the war and its aftermath. Norpoth et al.'s analysis, however, is based on an unweighted analysis of the available polling data. Does weighting the polls challenge their conclusions?

To answer that question, we (Caughey et al. 2017, pp. 22–24) compared the unweighted trend macropartisanship with weighted trends. As the dotted line in Fig. 1 indicates, several quarters before 1949 are observed during which the unweighted percentage of Democratic Party identifiers was matched by or even fell below the Republican percentage. However, when the data are weighted by phone ownership, professional status, and other consistently available auxiliary variables (dashed line labeled "comparable weights"), the estimated percentage of Democrats is 3–5 points higher, and it is 4–6 points higher when education (available beginning in 1943) is added to the auxiliary vector (solid line). In fact, the

[6] Some scholars who date the New Deal realignment to the 1930s attribute it solely to retrospective evaluations of the economy, as opposed to policy evaluations of the New Deal itself (Achen and Bartels 2016; but see Caughey et al. Forthcoming).

weighted estimates always are statistically distinguishable from 50%, and in many quarters are not statistically distinguishable from the 63% recorded in the 1952 ANES.[7] Especially considering that the "comparable" trend still probably underestimates Democratic ID in years before education weights are feasible, the Democrats' victory in 1948 looks less like a realignment than a local high point for a party that had enjoyed a clear majority since at least the mid-1930s.

Of course, the analysis does not definitively settle the question of what caused the New Deal realignment.[8] It does, however, illustrate how causal conclusions can hinge on (quantitative) descriptive inferences. Biased measurement of the outcome of interest can yield incorrect inferences about its causes. That is as true of the macrohistorical events and processes often of particular interest to APD scholars as it is of traditional CI designs focusing on the average effects of discrete treatments (e.g., Imai and Yamamoto 2010).

4 Example 2: estimating the effects of the Tennessee Valley Authority

We now turn to our second example: an analysis of the political effects of the Tennessee Valley Authority (TVA). The TVA was designed to modernize and develop the Tennessee River Valley, one of the poorest areas of the United States. Signed into law by President Franklin Delano Roosevelt in 1933, the TVA provided for a major increase in federal infrastructure spending in parts of Tennessee, Alabama, Mississippi, and Kentucky. Projects included electrification through the construction of hydroelectric dams, flood control projects, and major investments in roads and canals. Because of the massive size and importance of that intervention, as well as contemporaneous statements about its wide-reaching effects, scholars have studied the causal effects of the TVA across a variety of outcomes, both economic and political. But, while the TVA's footprint did expand dramatically after implementation, that factor alone does not provide evidence that TVA funding and projects were the cause of these changes. In this section of the paper, we discuss three identification strategies that we and other scholars have used to estimate the effects of the TVA on the region, and consider how those strategies fit into both CI and APD frameworks.

4.1 "Little TVAs" as counterfactuals

Two recent articles have examined the economic effects of the TVA: Kline and Moretti (2014); Kitchens (2014). Notably, those two papers come to somewhat contradictory conclusions based on quite different identification strategies. Both research designs depend heavily on assumptions about the process by which areas (in both cases, counties) were assigned to be "treated" by the TVA, and both papers marshall qualitative evidence to buttress their identification assumptions.

Kline and Moretti (2014) are interested in the long-run economic effects of the TVA, which (in brief) they find to be positive. Their identification strategy is closely tied to the counterfactual thinking at the heart of the RCM. It exploits the fact that, after the TVA was

[7] Note that the gap between the weighted and unweighted estimates narrows over time, which coincides with the gradual improvement in polling organizations's sampling techniques and their gradual transition to probability sampling after 1948.

[8] It also should be noted that Norpoth et al. (2013) present a good deal of micro-level evidence consistent with their theory.

implemented, legislation establishing similar development authorities in other parts of the country was introduced in Congress but never passed. The authors argue that those regions were economically similar to (though widely dispersed from) the Tennessee Valley, and that the legislation establishing the proposed authorities could easily have passed but did not for "exogenous political reasons" (Kline and Moretti 2014, p. 288). Thus, economic outcomes in counties covered by the proposed "little TVAs" provide plausible estimates for the counterfactual outcomes in TVA counties had the TVA never been created.

The paper's identification strategy obviously hinges on the claim that the failure of Congress to pass and implement the little TVAs was "exogenous", at least from an economic standpoint. The paper offers some evidence for that claim drawn from secondary sources, particularly in the form of *ex ante* judgments by contemporary observers that the (ultimately unsuccessful) legislation was likely to pass. Whether that evidence is persuasive depends in part on how credible it is to consider "political factors" unrelated to economic growth. If, for example, the TVA's location in the solidly Democratic South affected its growth prospects relative to proposed authorities in more Republican areas, then those areas would not provide reasonable counterfactuals. This highlights the fact that even many "natural experiments" cannot rest solely on claims about the treatment assignment, but also must invoke assumptions about the causes of the outcome—assumptions that always are context-specific.

Kline and Moretti's empirical approach also illustrates the limitations of the RCM for studying the development of a single country over time. As noted, Kline and Moretti's natural experimental design is based on a comparison of TVA and non-TVA counties. Ultimately, however, Kline and Moretti are at least as interested in whether the TVA boosted economic growth in the nation as a whole as they are in whether it helped the Tennessee Valley specifically. The problem is that place-based economic policies such as the TVA are likely to generate negative spillovers in non-targeted areas, making control counties a poor guide to what TVA areas would have experienced had the TVA never existed in the first place. Such "general equilibrium" effects violate the RCM's oft-neglected stable unit treatment value assumption (SUTVA; Rubin 1980). Kline and Moretti's interest in "agglomeration effects"—a form of economic path dependence (Pierson 2000, p. 255)—raises similar issues. To deal with thes difficulties, Kline and Moretti are forced to resort to a model-based estimation strategy based on a structural model of the US economy—an empirical approach more in tune with Pearl's SCM than with the RCM.

4.2 Distance from dams as an instrumental variable

In a study conducted around the same time as Kline and Moretti (2014), Kitchens (2014) too examines the TVA's effects on economic growth. Unlike Kline and Moretti, however, Kitchens focuses on one particular aspect of the TVA: its provision of subsidized electricity. Kitchens's study also is more narrowly focused—comparing TVA and non-TVA counties in the same state—and evinces deeper knowledge of the substantive and historical context of the TVA.

To motivate his approach, Kitchens highlights a variety of selection problems that make studying the treatment challenging. Those problems include the self-selection of counties and municipalities into TVA contracts, as well as the TVA's intentional targeting of "favorable test markets" wherein it was believed that the program was most likely to succeed (Kitchens 2014, pp. 390–2). To address those issues, Kitchens instruments for a county's use of TVA electricity with its distance from the nearest (electricity-producing) TVA dam. Using that instrument, Kitchens estimates that the economic effects of TVA electrification were indistinguishable from zero. To corroborate his conclusion that the TVA had little effect, Kitchens used records collected from various government and private archives to show that the TVA and nearby private utilities charged similar rates, suggesting that the TVA did not increase the productivity of its customers.

Kitchens's argument is considerably strengthened by the historical evidence he presents, which makes up a substantially larger proportion of the main text of his paper than that of Kline and Moretti (2014). Such evidence is crucial to the reader's understanding of the process by which counties received TVA power and to supporting his claim that dam sites were not influenced by the demand for electricity. Kitchens's archival material also provides (negative) evidence regarding the putative mechanisms of the TVA's effects, an example of the "causal process observations" advocated by Brady et al. (2006) and Dunning (2012, pp. 208–32). Such detailed, context-specific evidence would be difficult to collect had not Kitchens immersed himself in the historical workings of the TVA.

4.3 Matched panel design

In addition to its effects on economic outcomes, the TVA has long been suspected to have had important political effects. In particular, historians and other scholars of Southern politics have long documented that TVA areas tended to provide more support for pro-New Deal Democrats than other areas of the South, and that members of Congress (MCs) from TVA areas were more liberal than other Southern Democrats (Clapp 1963; Mayhew 1966; Rogers et al. 1994; Badger 2007). Many scholars have attributed that difference to the causal impact of the TVA itself, arguing that "[n]o government agency so strongly inspired southern demand for federal economic intervention as did the Tennessee Valley Authority" (Schulman 1994, p. 35). Those causal claims have not, however, been evaluated using the tools of modern causal inference. If corroborated, the TVAs effects on mass preferences and elite behavior would constitute an important case of policy feedback, a phenomenon that has been studied surprisingly rarely from a CI perspective (Campbell 2012, pp. 341–345).

We examine the political effects of the TVA in a working paper that provides yet another potential model for marrying CI and APD. We employ a modified version of a standard causal design: a matched panel (Heckman et al. 1997). Specifically, for each congressional term between 1931–1932 and 1961–1962, we first identified every House district that contained a cooperative or municipality serviced by TVA electrical power (Fig. 2). Then, we used genetic matching (Sekhon 2011) to match congressional districts covered by the TVA with observably comparable non-TVA districts elsewhere in the 13-state South (the former Confederacy plus Oklahoma and Kentucky). We then contrasted electoral results and congressional voting patterns in the treated and control districts in each term over that period. A distinctive feature of this design is that because district boundaries change over time (and not just decennially), it is not a true panel. We must therefore create new matches for each congressional term, always using the same set of pre-TVA characteristics as covariates.

This longitudinal design offers two important advantages. The first is the opportunity to conduct placebo tests (Rosenbaum 2002) of the TVA's "effects" before it was authorized in 1933. If matching is sufficient to balance the TVA and non-TVA districts on pre-treatment outcomes not included in the conditioning set, then our confidence that conditioning balances potential outcomes is strengthened. Second, the longitudinal design offers the opportunity to trace the effects of the TVA over time. That feature is especially valuable in this

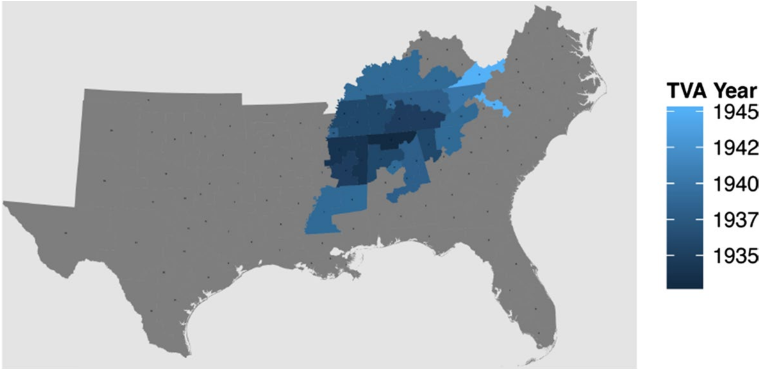
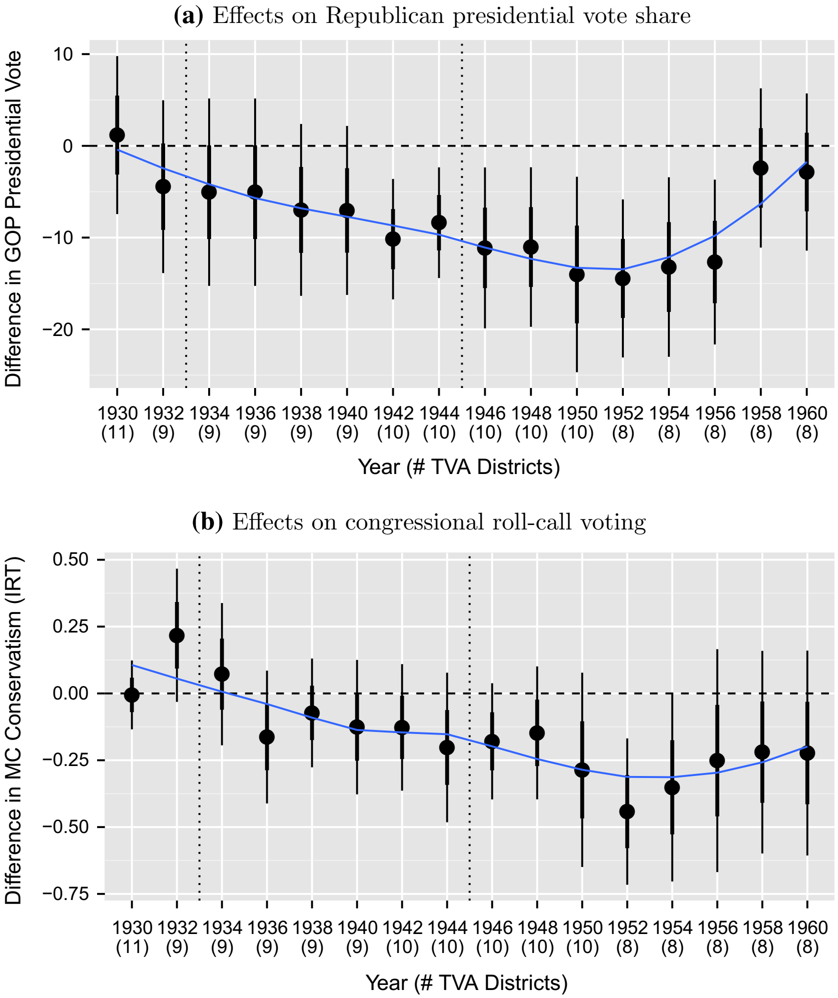


Fig. 2 House districts covered by the TVA in 1947–1948. Shading indicates the year that any part of the district was first served by the TVA

study because both the geographical reach of the TVA and the magnitude of its subsidies from the federal government varied markedly over time. The number of counties receiving TVA power increased gradually between 1933 and 1945, after which expansion largely ceased. Federal transfers to the TVA peaked in the first half of the 1950s, when per-household transfers reached 10% of average household income in the region. In the mid-1950s, congressional support for subsidizing the TVA dried up and, after 1959, the TVA essentially was self-financing. If the beneficiaries were sensitive to those funding changes, we would expect the political effects of the TVA to track their rise and decline.

In addition to suggesting a specific temporal pattern of effects, knowledge of the TVA and its historical context also informs our choice of outcome measures and the effects on them that we expect. First, although the TVA was the brainchild of Senator George Norris, a progressive Republican from Nebraska, it quickly became identified with the Democratic Party and, in particular, the party's left wing. Republicans, on the other hand, became the TVA's foremost opponents. That political context suggests the hypothesis that TVA districts would be relatively resistant to voting for Republican presidential candidates, whose vote shares in the South overall increased dramatically between the 1930s and the 1960s. This logic, however, does not necessarily extend to congressional elections, as the Democratic Party's lock on Southern House and (especially) Senate races remained virtually impenetrable through the 1950s (Black and Black 2002). Rather, most electoral competition for Southern congressional seats continued to take place in Democratic Party primaries, between left-leaning and right-leaning members of the same party (Caughey 2018). That observation motivates our second main hypothesis that representatives from TVA districts would be less likely to support conservative economic positions. We operationalize that outcome variable by estimating a dynamic item response theory (IRT) model on House roll calls related to social welfare or government regulation, which yields estimates of members' economic conservatism.

Figure 3 illustrates our empirical results. Both panels compare TVA districts with controls matched on the basis of demographic and economic characteristics measured before the authorization of the TVA. In the analysis, political characteristics were not used


(a) Effects on Republican presidential vote share

(b) Effects on congressional roll-call voting

Fig. 3 Estimated effects of the TVA on Republican presidential vote (top) and conservative roll-call voting (bottom). In each election year, districts are matched on pre-1933 demographic and economic (but not political) characteristics. Dotted lines demarcate the expansion period of the TVA

to create the matches so as to preserve our ability to conduct placebo tests for 1930 and 1932.[9] (Adding political attributes to the covariate set further shrinks the pretreatment differences while increasing the precision of the estimated effects.) TVA districts differed little from non-TVA ones before the program's implementation, but for about a decade after implementation exhibited markedly lower Republican vote shares and congressional conservatism. The evidence suggests that the TVA did in fact increase support for New Deal liberalism in the areas it covered, relative to non-covered areas.

[9] The years refer to election years; members' conservatism is measured in the following congressional term (e.g., 1933–1934 for 1932).

Our study of the TVA, together with Kitchens (2014) and Kline and Moretti (2014), illustrates the power as well as some of the limitations of combining causal inference with historical and developmental inquiry. On one hand, CI clearly provides valuable evidence about why the South converged economically with the rest of the United States, why some parts of the South turned against the New Deal and Democratic Party more sharply then elsewhere, and other important questions of economic and political development. Moreover, a rich understanding of the historical context was critical to the hypotheses and research designs of the three studies summarized herein. Merely identifying the TVA as a policy intervention that might provide leverage for CI-style analysis would not get a researcher very far; qualitative, historical research was necessary for evaluating the plausibility of potential counterfactuals, identifying appropriate outcomes of interest, and ruling out alternative explanations. In those respects, then, causal inference and APD fit well together.

On the other hand, CI cannot always provide clear answers to political developmental questions, for at least two reasons. First, as is true of all applications of CI, the causal conclusions drawn from a given study are only as good as the research design and identification strategy it employs. Although identification assumptions can be supported with historical evidence, they rarely can be demonstrated conclusively. Consequently, as the examples of Kitchens (2014) and Kline and Moretti (2014) show, two different but equally plausible research designs can lead to divergent conclusions, for reasons ranging from heterogeneous effects to faulty assumptions. Second, even when credible, the estimates from a CI analysis often do not answer questions of substantive interest. CI's focus on "effects of causes" rather than "causes of effects" often precludes it from offering fully satisfying *explanations* for observed outcomes (e.g., why did white Southerners' turn against the New Deal?).[10] Further, as Kline and Moretti's use of structural models highlights, the nonparametric orientation of the RCM can make it difficult if not impossible to tackle questions about general equilibria and other aggregate effects (e.g., what were the economic benefits of the TVA to the United States as a whole?).

5 Conclusion

In this article, we have discussed how we have thought about causality in our own work as well as some broader considerations on how the methods and approaches of American political development (APD) and causal inference (CI) can be brought together. In many ways, APD and CI are quite different worlds, and the kinds of causality they primarily are interested in are not always easy to combine. CI typically is most interested in sharp, manipulable causes that act upon many units that do not interfere or depend upon one another—all features that can be challenging to integrate with APD approaches. Ultimately, bringing the two fields together should mean more than searching through the historical record for natural experiments. APD scholars can and should pursue CI-style analyses when their data are appropriate for those designs, but even when inferences in the CI/Rubin causal model mold can be made using historical data, the causal estimates often don't capture everything we might want to know about the development of American politics over

[10] By (good) explanation, we mean a statement of valid premises from which the outcome of interest follows necessarily or with high probability, and that specifies the mechanism by which the premise entails the outcome.

time. Instead, many interesting historical arguments are best made using a combination of quantitative and qualitative evidence, including both CI-style causal estimates and descriptive evidence. We argue that, where possible and appropriate to their research questions, scholars should embrace a less narrow conception of causality that allows for bringing to bear as evidence both precise causal estimates of particular treatments as well as considering how context and timing can change the effect of treatments and the ways in which treatments might interact at specific moments in history.

References

Achen, C. H., & Bartels, L. M. (2016). *Democracy for realists: Why elections do not produce responsive government*. Princeton, NJ: Princeton University Press.

Badger, A. J. (2007). Whatever happened to Roosevelt's new generation of southerners? In *New Deal/New South* (pp. 58–71). University of Arkansas Press, Fayetteville.

Berinsky, A. J. (2006). American public opinion in the 1930s and 1940s: The analysis of quota-controlled sample survey data. *Public Opinion Quarterly*, *70*(4), 499–529.

Berinsky, A . J., Powell, E . N., Schickler, E., & Yohai, I . B. (2011). Revisiting public opinion in the 1930s and 1940s. *PS: Political Science & Politics*, *44*(3), 515–520.

Black, E., & Black, M. (2002). *The rise of southern republicans*. Cambridge, MA: Belknap Press.

Brady, H. E. (2009). Causation and explanation in social science. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology*. New York: Oxford University Press.

Brady, H. E., Collier, D., & Seawright, J. (2006). Toward a pluralistic vision of methodology. *Political Analysis*, *14*(3), 353–368.

Burnham, W. D. (1967). Party systems and the political process. In W. N. Chambers & W. D. Burnham (Eds.), *The American party systems: Stages of political development* (pp. 277–307). New York: Oxford University Press.

Campbell, A. L. (2012). Policy makes mass politics. *Annual Review of Political Science*, *15*, 333–351.

Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Belmont, CA: Wadsworth.

Caughey, D. (2018). *The unsolid south: Mass politics and national representation in a one-party enclave*. Princeton, NJ: Princeton University Press.

Caughey, D., Berinsky, A. J., Chatfield, S., Hartman, E., Schickler, E., & Sekhon, J. J. (2017). Population estimation and calibration weighting for nonresponse and sampling bias: An application to quota-sampled opinion polls (pp. 1936–1952). Unpublished manuscript.

Caughey, D., & Chatfield, S. (2016). Creating a constituency for New Deal liberalism: The policy feedback effects of the Tennessee Valley Authority. Paper presented at the APSA annual meeting, Philadelphia, PA.

Caughey, D., Dougal, M. C., & Schickler, E. (Forthcoming). Policy and performance in the New Deal realignment: Evidence from old data and new methods. *Journal of Politics*.

Caughey, D., & Wang, M. (2019). Dynamic ecological inference for time-varying population distributions based on sparse, irregular, and noisy marginal data. *Political Analysis*, *27*, 388–396.

Clapp, C. L. (1963). *The congressman: His work as he sees it*. Garden City, NY: Anchor Books.

Converse, J. M. (1987). *Survey research in the United States: Roots and emergence*. Berkeley: University of California Press.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382.

Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach*. New York: Cambridge University Press.

Hacker, J. S. (2002). *The divided welfare state: The battle over public and private social benefits in the United States*. New York: Cambridge University Press.

Hartman, E., Grieve, R., Ramsahai, R., & Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(3), 757–778.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*(4), 605–654.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(296), 945–960.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765–789.

Imai, K., & Yamamoto, T. (2010). Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science*, *54*(2), 543–560.

Key, V. O, Jr. (1964). *Politics, parties and pressure groups*. New York: Crowell.

Kitchens, C. (2014). The role of publicly provided electricity in economic development: The experience of the Tennessee Valley Authority, 1929–1955. *Journal of Economic History*, *74*(2), 389–419.

Kline, P., & Moretti, E. (2014). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee Valley Authority. *Quarterly Journal of Economics*, *129*(1), 275–331.

Ladd, E. C., & Hadley, C. D. (1975). *Transformations of the American party system: Political coalitions from the New Deal to the 1970s*. New York: Norton.

March, J. G., & Olsen, J. P. (1984). The new institutionalism: Organizational factors in political life. *American Political Science Review*, *78*(3), 734–749.

Mayhew, D. R. (1966). *Party loyalty among congressmen: The difference between democrats and republicans, 1947–1962*. Cambridge, MA: Harvard University Press.

Mayhew, D. R. (2002). *Electoral realignments: A critique of an American genre*. New Haven, CT: Yale University Press.

Mettler, S., & Valelly, R. (2016). Introduction: The distinctiveness and necessity of American Political Development. In R. Valelly, S. Mettler, & R. Lieberman (Eds.), *The Oxford handbook of American political development*. Oxford University Press.

Morgan, S . L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). New York: Cambridge University Press.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Roiniczych, Tom X* (pp. 1–51).

Norpoth, H., Sidman, A. H., & Suong, C. H. (2013). Polls and elections: The New Deal realignment in real time. *Presidential Studies Quarterly*, *43*(1), 146–166.

Orren, K., & Skowronek, S. (2004). *The search for American political development*. New York: Cambridge University Press.

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. New York: Oxford University Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference* (1st ed.). New York: Cambridge University Press.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146.

Pierson, P. (1993). When effect becomes cause: Policy feedback and political change. *World Politics*, *45*(4), 595–628.

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American Political Science Review*, *94*(2), 251–267.

Rogers, W. W., Ward, R. D., Atkins, L. R., & Flynt, W. (1994). *Alabama: The history of a deep south state*. Tuscaloosa: University of Alabama Press.

Rosenbaum, P . R. (2002). *Observational studies* (2nd ed.). New York: Springer.

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data: The Fisher randomization test," by D. Basu. *Journal of the American Statistical Association*, *75*(371), 591–593.

Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, *15*(1), 38–46.

Särndal, C.-E., & Lundstrom, S. (2005). *Estimation in surveys with nonresponse*. Hoboken, NJ: Wiley.

Schickler, E. (2001). *Disjointed pluralism: Institutional innovation and the development of the U.S. Congress*. Princeton, NJ: Princeton University Press.

Schulman, B. J. (1994). *From Cotton Belt to Sunbelt: Federal policy, economic development, and the transformation of the South, 1938–1980*. Durham, NC: Duke University Press.

Sekhon, J. S. (2008). The Neyman–Rubin model of causal inference and estimation via matching methods. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 271–299). New York: Oxford University Press.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The `Matching` package for R. *Journal of Statistical Software*, *42*(7), 1–52.

Sen, M., & Wasow, O. (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, *19*(1), 499–522.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*, *15*(1), 3–17.

Sundquist, J. L. (1983). *Dynamics of the party system: Alignment and realignment of political parties in the United States* (Revised ed.). Washington, DC: Brookings.

Thelen, K. (1999). Historical institutionalism in comparative politics. *Annual Review of Political Science*, *2*, 369–404.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.